Combining Proper Name-Coreference with Conditional Random Fields for Semi-supervised Named Entity Recognition in Vietnamese Text

Rathany Chan Sam, Huong Thanh Le,
Thuy Thanh Nguyen, and Thien Huu Nguyen

Hanoi University of Science and Technology,
1 DaiCoViet street, Hanoi, Vietnam
rathany_cam@yahoo.com,
{huonglt,thuynt}@it-hut.edu.vn,
nguyenhuuthien88bk@yahoo.com

Abstract. Named entity recognition (NER) is the process of seeking to locate atomic elements in text into predefined categories such as the names of persons, organizations and locations. Most existing NER systems are based on supervised learning. This method often requires a large amount of labelled training data, which is very time-consuming to build. To solve this problem, we introduce a semi-supervised learning method for recognizing named entities in Vietnamese text by combining proper name coreference, named-ambiguity heuristics with a powerful sequential learning model, Conditional Random Fields. Our approach inherits the idea of Liao and Veeramachaneni [6] and expands it by using proper name coreference. Starting by training the model using a small data set that is annotated manually, the learning model extracts high confident named entities and finds low confident ones by using proper name coreference rules. The low confident named entities are put in the training set to learn new context features. The F-scores of the system for extracting "Person", "Location" and "Organization" entities are 83.36%, 69.53% and 65.71% when applying heuristics proposed by Liao and Veeramachaneni. Those values when using our proposed heuristics are 93.13%, 88.15% and 79.35%, respectively. It shows that our method is good in increasing the system accuracy.

Keywords: information extraction, named entity extraction, entity coreference, semi-supervised learning, CRFs.

1 Introduction

Named Entity Recognition is a subtask of information extraction. Its purpose is to identify and classify certain proper nouns into some predefined target entity classes such as person, organization, location and temporal expressions.

Much previous work on NER followed the supervised learning approach [2], [3], [9], [12], [15] which requires a large hand-annotated corpus. Such approaches

J.Z. Huang, L. Cao, and J. Srivastava (Eds.): PAKDD 2011, Part I, LNAI 6634, pp. 512–524, 2011.

can achieve good performances. However, annotating such a corpus requires a lot of human effort. This problem can be solved by using a sequence-based semi-supervised method that trains a classification model on an initial set of labelled data, makes predictions on a separate set of unlabelled data, and then iteratively attempts to create an improved model using predictions of the previously generated model (plus the original labelled data). Based on this method, we propose a semi-supervised learning method for recognizing named entities in Vietnamese text by combining proper name coreference, named-ambiguity heuristics with a powerful sequential learning model, Conditional Random Fields (CRFs). Our approach inherits the idea of Liao and Veeramachaneni [6] and expands it by using proper name coreference. Starting by training the model using a small data set that is tagged manually, the learning model extracts high confident named entities with and finds low confident NEs by using proper name coreference rules. The low confident NEs are put in the training data set to learn new context features.

Example 1.

(a) *Hôm nay có mưa lớn ở* ***Thành phố Hồ Chí Minh*** /It rains heavily in Hochiminh city today.

(b) *Chính phủ đang tìm giải pháp chống tắc đường ở* ***TP HCM*** /The government is finding a method to solve traffic jam in Hochiminh city.

In Example 1, both "*Thành phố Hồ Chí Minh*/Hochiminh city" and "*TP HCM*" are Location entities and refer to one location. However, the system can only find one Location entity with high confident score, which is "*Thành phố Hồ Chí Minh*/Hochiminh city". The phrase "*TP HCM*" is not recognized as a Location entity by the system since the confidence score of this phrase is smaller than the threshold. Based on the coreferent rules, the system discovers that "*Thành phố Hồ Chí Minh*/Hochiminh city" and "*TP HCM*" refer to the same location. From that point of view, "*TP HCM*" is considered as a low confidence part of "*Thành phố Hồ Chí Minh*/Hochiminh city". "*TP HCM*" is then forced to be a Location entity. It is put in the training set to relearn the new feature context.

In addition, based on our empirical study, several named entity (NE) rules are manually added to the system, in order to create new training data from unlabelled text.

The rest of this paper is organized as follows. Section 2 introduces recent studies on semi-supervised NER methods and works that inspire our research. Section 3 briefly introduces include CRF and the training and inference of the CRF. Section 4 discusses the semi-supervised NER problem for Vietnamese text and our solution to this problem. Section 5 analyzes our experimental results. Finally, our conclusions and future work are given in Section 6.

2 Related Works

The term "semi-supervised" (or "weakly supervised") is relatively recent. The main technique for semi-supervised learning is called "bootstrapping" and involves a small degree of supervision for starting the learning process.